
Automated preparation of DNA sequences for publication

Marvin B. Shapiro and Periannan Senapathy

Laboratory of Statistical and Mathematical Methodology, Division of Computer Research and Technology, National Institutes of Health, Bethesda, MD 20205, USA

Received 15 April 1985

ABSTRACT

A computer program which draws DNA sequences is described. A simple method is used which enables the user to highlight or annotate specific parts of a sequence. The sizes of the characters in the sequence to be drawn are specified by the user. In addition, vertical spacing between lines and horizontal spacing between characters can be specified. Sequences can be prepared and high quality output produced on a plotter in a short period of time, making the program advantageous to use over typing, computer printing, or preparation by a graphics department.

INTRODUCTION

Technical progress in the isolation and sequence analysis of DNA has greatly accelerated the rate of accumulation of new sequences. Of late, published sequences include complementary data bearing on genome organization, transcription, RNA processing, translation and the evolutionary comparisons of different sequences. Simple listing of DNA sequences from a typewriter or a computer file is not sufficient to handle the preparation of sequences for publication. The computer can be used to provide high quality drawings of DNA sequences in a short amount of time. Other advantages to using the computer include the following: (i) A wide selection of characters is available. (ii) Spacing of characters can be specified. (iii) Data that is already in computer form does not have to be retyped. (iv) Special characters not available on a standard typewriter or computer printer can be used (e.g. arrows, italics, lines). (v) Emphasis of characters (e.g. using underlining, boxes, colors, and different character sizes) is easily done. A computer program to take care of the various requirements in the publication of DNA sequences would be extremely helpful for workers in this field. We report here such a program, DNADRAW, that draws publication quality pictures of DNA sequences, which include notes and

The DNADRAW program, and the auxiliary programs for displaying and plotting, are available from the authors at no charge. Send a self-addressed mailing label and a blank tape. In order to use this system a DECsystem-10 computer and CalComp plotter are required.

Table 1. Different options of the DNADRAW program

Option	Example	Meaning or Symbol
underlining	ACGT --	CG is underlined. Line ends with \$.
box	ACGT bb	CG is enclosed in a box.
open box	ACG bo	Right side of box around CG is open.
open box	CGT ob	Left side of box around CG is open.
italics	ACGT ii	CG is drawn in italics.
greeks	a cell g	a is drawn as alpha.
small characters	ACGT ss	CG is drawn smaller.
large characters	ACGT ll	CG is drawn larger.
not equal spacing	ACGT ppp	ACG are proportionally spaced.
centering	ACGT ccc	ACG are proportionally spaced and centered over the A.
right justifying	ACGT rrr	Proportional spacing, right justified at the T.
coloring	AGCGT d e n	A,C,T are colored red, blue, and green, respectively.
subscripting	a1 j	1 is a subscript.
superscripting	b2 k	2 is a superscript.
polygon	ACGT u u	Two corners of polygon u are at the top left of A and G.
partial blank	.5\$	Half a blank line.
extra space	2.5\$	2.5 blank lines.

The second line of each pair of example lines in the table (except the last two lines) end with "\$", and are used to modify the line above it. The modifying characters "a", "b", "c", "d", "e", "g", "i", "j", "k", "l", "n", "o", "p", "r" and "s" are used as shown in the table. Lines ending with "\$" can also be used for vertical spacing, as shown in the last two lines of the table.

Table 2. Special characters drawn by the DNADRAW program

Character	Drawn as
@	↓, shifted left a half space
#	↓
%	↑, shifted left a half space
^	↑
&	→
\	←
(underline) _	horizontal line

The seven keyboard characters listed in Table 2 are drawn as shown at the right. All other characters on the keyboard are drawn as they are.

special characters for pointing out structural and functional aspects of the sequences at desired places.

THE DNADRAW PROGRAM

Input to the program consists of files of DNA and protein sequence, and other descriptive information. The user answers questions on how the files are to be handled. The user's choice of display and plot options determines whether the sequence is displayed on a graphics screen and whether a hard copy is drawn with a plotter.

The actual drawing of the characters by DNADRAW is done by another program, GRAPH (1), written for general purpose drawing of graphs. GRAPH has all the capabilities for character and line drawing, for equal spacing, and for drawing different fonts and special characters.

If the raw DNA data is from a data bank or is in the form of a long string of characters, then a preliminary step is necessary to break up the data into individual lines that are formatted. This is done using the FORMAT option of the DNADRAW program, as illustrated below in Example 3.

The file of DNA information used as input to DNA is prepared as a series of lines containing sequence information, with special symbols added to get underlining, boxed letters, italics, colors, size changes, proportionally spaced letters and spacing (see Table 1). Other symbols are used to get arrows and horizontal lines (see Table 2). Underlining and the other features described in Table 1 are obtained by adding one or more extra lines, ending with a \$, below the line to be modified. These lines are used only for modifying previous lines, and do not appear as part of the output. The modifying lines are added using a text editor.

Keyboard characters @, #, %, ^, &, and \ are used to obtain arrows, as shown in Table 2, and _ (underline) is used for indicating sequence homologies. The shifted arrows are used to point between nucleotides.

A

```

Sma I @          Cat box      #
ppppp           cccccc$
GCGTGTACCCGGGGCTTCTTGACCAATTACCTCTGACCTGTCATACCCC
-----
                    bbbbb$

Promoter
ccccccc$
GCGCTATAAATGAGCTCAGAGTAGCGTAGCTACTCCACCGGGAGGTAA
bbbbbb$

*      |_ Exon I&
                    ppppppp$
GCTGGGATCGTCACCGATGCTTCTCGCTCACGAGGGGAACGTGGCTATCT
MetLeuLeuAlaGlnGluGlyAsnValAlaIleS

                    donor splice
                    cccccccccc$
CCATTAGACTGAAAGCCCTGTGGGAGGTAAGTGAAGTGCACGCTCGAT
-----$
erlleArgLeuLysAlaProValGlyG|_ Intron I&
                    ppppppppp$

ACCCCTTGCTGCTTAACGAGCAAGCTGTAGCTAGCGTAGTTACGATCGCTG

CTGACTCGTGGCCTGAATATCCTTGTGATGCTCTATTTTCGGATCGTG
bbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbbb$
                    Enhancer sequence
                    cccccccccccccccc$

|_ Exon II&
pppppppp$
acceptor splice
ccccccccccccccc$
GCTGACCCAGGTATTAGACTGAAAGCCCTGTGGGAGACAAGTGAAGTCA
-----$
lylleArgLeuLysAlaProValGlyAspAsnEnd
iiiiiii$

TGTTCACCGCGCGTCACGGTTCAATAAAACCTCTACACTACGGTTAACTG
-----$
Hpa I$
ppppp$

-1$
Polyadenylation signal
ccccccccccccccccccccccc$

TACGAGTITTTTTTTTTTCTACTATATATATATATATATATGCGATC
-----
iiiiiiiiiiiiiiiiiii$
```

B

```

Sma I |          Cat box      |
GCGTGTACCCGGGGCTTCTTGACCAATTACCTCTGACCTGTCATACCCC
GCGCTATAAATGAGCTCAGAGTAGCGTAGCTACTCCACCGGGAGGTAA
Promoter
GCTGGGATCGTCACCGATGCTTCTCGCTCACGAGGGGAACGTGGCTATCT
MetLeuLeuAlaGlnGluGlyAsnValAlaIleS
Exon I
CCATTAGACTGAAAGCCCTGTGGGAGGTAAGTGAAGTGCACGCTCGAT
erlleArgLeuLysAlaProValGlyG|_ Intron I
ACCCCTTGCTGCTTAACGAGCAAGCTGTAGCTAGCGTAGTTACGATCGCTG
CTGACTCGTGGCCTGAATATCCTTGTGATGCTCTATTTTCGGATCGTG
Enhancer sequence
GCTGACCCAGGTATTAGACTGAAAGCCCTGTGGGAGACAAGTGAAGTCA
acceptor splice
TGTTCACCGCGCGTCACGGTTCAATAAAACCTCTACACTACGGTTAACTG
Polyadenylation signal
TACGAGTITTTTTTTTTTCTACTATATATATATATATATATGCGATC
Hpa I
```

Figure 1. A) An example of a DNA sequence file, as input to the DNADRAW program. In addition to the DNA and protein sequence lines, this file has additional lines, containing the names of the restriction enzymes, CAT box, and symbols for arrows, etc. at appropriate places. Underlines, boxes, and proportional spacing are indicated by a modifier line ending in \$ with symbols as shown in Table 1. B) The plotter output.

Example 1: A sample input to the DNADRAW program and the plotter output.

Example 1 shows, for a short DNA sequence, the input to the DNADRAW program (Figure 1A) and the final plotted output (Figure 1B). Note how the "\$" lines in Figure 1A modify the lines above to give the output shown in Figure 1B, and how the annotations (Sma I, etc.) are proportionally spaced and sometimes centered.

Example 2: Drawing sequence homologies.

Frequently, special symbols such as ":" or horizontal lines are used to indicate sequence homologies. Example 2 illustrates the use of the "_" (underline) symbol to indicate sequence homologies. Figure 2A shows the input and Figure 2B the sequence drawn.

A CATTAGCCCGAC _C___GGT___ GGTACTCAT _A___C_	B CATTAGCCCGAC -C-----GGT- GGTACTCAT -A-----C-
---	---

Figure 2. A) The homology between two different DNA sequences is indicated by underline characters in the input file. B) In the output file horizontal lines are drawn.

Example 3: Formatting and drawing.

In this example the unformatted DNA and protein files, PRODAT and DNADAT, shown in Figure 3A are modified and used as input to DNADRAW, to produce the plotter output shown in Figure 3D. The two steps taken to produce that figure are given in detail. Default answers to the program prompts are in square brackets ([]), and user responses are underlined.

Step 1. The FORMAT option of the DNADRAW program is used to format data files PRODAT and DNADAT into 50 column lines, add a count to the right of the DNA lines, and merge the two files into an output file named FIG3B. The amino acid sequence starts at the 29th nucleotide, which the user indicates. DNADRAW inserts a blank line into the output file after each pair of PRODAT and DNADAT lines. The program prompts and user responses are the following:

```
.DNADRAW
FORMAT OPTION ?[N] Y
# INPUT FILES = ? 2
NAME OF INPUT FILE 1 = ? PRODAT
STARTING POSITION FOR FILE 1 = [1] 29
NAME OF INPUT FILE 2 = ? DNADAT
STARTING POSITION FOR FILE 2 = [1] _
# CHARACTERS PER LINE = ? 50
ADD A COUNT TO WHICH FILE ? [NONE] 2
  TO THE RIGHT (R) OR LEFT (L) ? [R] _
NAME OF OUTPUT FILE = ? FIG3B
```

Step 2. File FIG3B is edited to get file FIG3C, shown in Figure 3C. Italics, a box, underlining, and a first line containing column numbers are added. Note the use of two consecutive \$ lines to get italics AND underlining.

A

```

PRODAT      MetGlnCysHisHisThrCysThrHis
            ThrArgAlaProThrGlnThrSerAlaProPro
            HisProSerSerLeuSerArgGlyEnd

            AGGAGGAGTCATTGCAAAGAAAAATCGAGATGCAATGCCACCACACGTGCA
            CACACGTGCACCCACACACGTGCACCCACACAAACC
DNADAT      TCTGCTCCTCCCCACCCCTCTTCACTTTCCAGAGGATAAGGCAAAATGTGAGCGCACCCCA
            GGTGCTTGTGTTTGACCCATCAGAAGCAGAGATCAACAAAACCCAGAAGGCCACGCTCGT
            TACCCTGACCACGTGGAGCTGAG

```

B

```

                                MetGlnCysHisHisThrCysT
AGGAGGAGTCATTGCAAAGAAAAATCGAGATGCAATGCCACCACACGTGCA    50

hrHisThrArgAlaProThrGlnThrSerAlaProProHisProSerSer
CCCACACACGTGCACCCACACAAACCTCTGCTCCTCCCCACCCCTCTTCA    100

LeuSerArgGlyEnd
CTTTCCAGAGGATAAGGCAAAATGTGAGCGCACCCAGGTCGTTGTGTTTG    150

ACCCATCAGAAGCAGAGATCAACAAAACCCAGAAGGCCACGCTCGTTACC    200

CTGACCACGTGGAGCTGAG    219

```

C

```

          10          20          30          40          50
                                MetGlnCysHisHisThrCysT
                                iiiiiiiiiiiiiiiiiiiii$
AGGAGGAGTCATTGCAAAGAAAAATCGAGATGCAATGCCACCACACGTGCA    50
                                ---$
                                -----$

hrHisThrArgAlaProThrGlnThrSerAlaProProHisProSerSer
                                u          u$
CCCACACACGTGCACCCACACAAACCTCTGCTCCTCCCCACCCCTCTTCA    100
- ----- bbbbbbb$

                                u          u$
LeuSerArgGlyEnd
CTTTCCAGAGGATAAGGCAAAATGTGAGCGCACCCAGGTCGTTGTGTTTG    150

ACCCATCAGAAGCAGAGATCAACAAAACCCAGAAGGCCACGCTCGTTACC    200

CTGACCACGTGGAGCTGAG    219
ss      111$

```

D

```

          10          20          30          40          50
                                MetGlnCysHisHisThrCysT
AGGAGGAGTCATTGCAAAGAAAAATCGAGATGCAATGCCACCACACGTGCA    50
                                MetGlnCysHisHisThrCysT
hrHisThrArgAlaProThrGlnThrSerAlaProProHisProSerSer
GCCACACACGTGCACCCACACAAACCTCTGCTCCTCCCCACCCCTCTTCA    100
                                MetGlnCysHisHisThrCysT
LeuSerArgGlyEnd
CTTTCCAGAGGATAAGGCAAAATGTGAGCGCACCCAGGTCGTTGTGTTTG    150

ACCCATCAGAAGCAGAGATCAACAAAACCCAGAAGGCCACGCTCGTTACC    200

CTGACCACGTGGAGCTGAG    219

```

Figure 3. A complete example, showing how formatting, merging, and editing of individual files containing DNA and protein sequences is done. A) the files PRODAT and DNADAT of amino acid and DNA sequences contain lines of unequal lengths. The FORMAT option of the program merges and formats these files, also adding a count as instructed by the user. B) Sequences formatted by the FORMAT option of the program. The amino acid sequence starts at position 29 of the DNA sequence as specified by the user. C) The file is edited by adding modifier lines, which make the specified characters *italics*, boxed, small or large. D) The final output plotted by a Calcomp plotter.

Step 3. DNADRAW is called again to draw the sequence. The dialog is shown below. There are 55 characters on the longest lines (50 + 5 for the count). The default width of 7.9" for each line based on a ratio of 7 characters per inch for good quality letters, is used. A character width factor of 1.2 is used to provide more space between the characters drawn. The space allowed for each character is equal to

WIDTH OF EACH CHARACTER (IN INCHES)

CHARACTERS IN LONGEST LINE

If the default width factor of 1 is used, the space allotted will be the width of "W", the widest character in the alphabet. If a width factor of say 1.5 is used the "W" will fill only $2/3 (= 1/1.5)$ the space allotted, i.e. a smaller character will be drawn to fit into the same space. A factor greater than 1 may be necessary when using italics or large characters because they can be wider than "W".

An added feature of the program is the ability to control the vertical spacing between lines, the default, 1, giving the standard spacing seen in most publications. A factor of less than 1 is used to bring lines closer together, and greater than 1 to stretch the picture out. A blank line is the same as a 1\$ line (see Table 1), i.e. vertical space for one line is allotted. If a blank line is replaced with say 2.5\$, two and a half lines of vertical space are provided. These kinds of horizontal and vertical spacing controls are not available with a typewriter or computer printer.

The dialog for step 3 given below specifies that the sequence is to be displayed. After the sequence is displayed and seen to be correct, plotter output is specified. The final output is shown in Figure 3D.

.DNADRAW

FORMAT OPTION ? [N] _

NAME OF INPUT FILE = ? FIG3C

CHARACTERS IN LONGEST LINE = ? 55

WIDTH OF EACH LINE (IN INCHES) = ? [7.9] _

CHARACTER WIDTH FACTOR = ? [1] 1.2

VERTICAL SPACING FACTOR = ? [1] _

DISPLAY THE SEQUENCE ? [Y] _

(The sequence is displayed)

PLOT THE SEQUENCE ? [Y] _

Example 4: Illustration of rectangular polygons.

Figure 4 shows the output for a sequence containing 13 rectangular

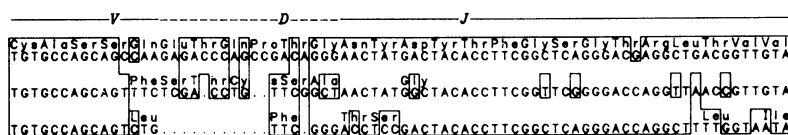


Figure 4. An illustration of the use of polygons to indicate homologies of sequences.

polygons. They are specified by using a different letter for each polygon, and by placing the letters on \$ lines to indicate the corners of the polygons, as shown in Figure 3C. Any letter not already used in Table 1 can be used (i.e. a,f,h,m,q, and t-z).

DISCUSSION

It is time-consuming and expensive to prepare DNA sequences for publication through a graphics department. While the use of a typewriter or computer printer is less expensive, they allow no flexibility in terms of the size and spacing of characters, and their characters are often inferior in quality.

There are a number of features of the DNADRAW program which overcome these deficiencies. First, the quality of the output is excellent, and in addition many special characters (arrows, italics, etc.) can be drawn by the computer. The requirement of equal spacing of characters, to give alignment of sequences, is provided by the program, whereas this is difficult to do with many graphic typesetting machines.

The preparation of a sequence file used as input for the DNADRAW program is relatively simple. The additional modifying lines (ending in "\$") line up with the letters modified, making it easy to see the result and to check the entire sequence before drawing it. A file of sequence data with 1000 nucleotides and 1000 other characters for amino acids, numbering, arrows, etc. should take less than a half hour to prepare (assuming the sequence data start out in computer form). Displaying the 2000 characters takes about 5 minutes, but this step can be skipped, and the plot takes about 9 minutes. Thus, it should be possible for a sequence containing 20,000 characters to be prepared and completed for publication in less than half a day with this program. Projecting further, it would be relatively inexpensive and practical to use a program such as this to produce an entire book of sequences.

The ideas presented here should help in the development of similar programs for handling the huge volume of DNA sequence information now entering the literature.

The DNADRAW program is written in the SAIL language, and runs on a DECsystem-10 computer.

ACKNOWLEDGEMENTS

We thank Dr. Rose Mage for providing the sequence used in example 4 and the many colleagues at NIH who offered suggestions, which resulted in a greatly improved program.

REFERENCES

1. Shapiro, M.B. (1984) 'Graph: A program for drawing graphs', National Institutes of Health - Division of Computer Research and Technology Publication.